Exact analysis of a paired sibling study

I. H. Dinwoodie
Department of Mathematics, Tulane University

Brenda MacGibbon
Department of Mathematics and Statistics, Université du Québec à Montréal

August 21 2001

Abstract. A data set on categories of congenital heart malformations for sibling pairs of Fraser and Hunter (1975) is analyzed exactly for quasi-independence with Monte Carlo methods. Exact p-values are computed for a test of parameter significance and a test of goodness-of-fit which contradict the model of quasi-independence and confirm an earlier analysis of MacGibbon (1983).

Key words. exact conditional test, exponential family, Monte Carlo, quasi-independence, structural zeros

1. Introduction.

Exact methods for categorical data have recently been the subject of renewed statistical interest because contingency tables are arising in application areas such as genetics which have integer entries of counts small enough in some cells to cause doubt about the validity of multivariate normal approximations. With some entries near zero, there is concern about the validity of normal approximations whose accuracy gets worse for multinomial probabilities near the boundary of the probability simplex. This is the same phenomenon as the well-known fact that the normal approximation to the binomial is worse (for fixed sample size n) when the success probability parameter p is near 0 or $q = 1 - p$ is near 0. On the other hand, the tables in these applications have entries large enough in other cells to make enumeration difficult.

In this paper we analyze exactly such a data set of Fraser and Hunter (1975). The results confirm earlier conclusions of Fraser and Hunter, and MacGibbon (1983). The example is of interest for two reasons. First, the conclusions should be of interest to researchers in biology since they confirm the original discovery of Fraser and Hunter concerning the coincidence of Tetralogy of Fallot (ToF) and Pulmonary Stenosis (PS). Whereas p-values for exact tests can be larger than approximate p-values because of the discrete data, the significance in this case is strongly confirmed. Second, the example illustrates a slight extension to nontriangular tables of an exact simulation method for the hypergeometric distribution on triangular tables (with fixed row and column sums) which should be of interest to statisticians.

Exact inference is the enterprise of computing p-values exactly and does not use asymptotic probability approximations for the distribution of a test statistic. Typically it uses distributions conditional on a sufficient statistic so the result is uniform over all distributions in the null family. The survey papers of Agresti (1992, 1999) and the paper of Diaconis and Sturmfels (1998) give many references to recent applications and methods of exact conditional tests.

The computations are of probabilities of events in nonnegative integer lattice points (level sets for sufficient statistics) with respect to a multivariate hypergeometric distribution. Exact computational methods fall into two groups: complete enumeration and Monte Carlo methods. Complete enumeration is represented by work of Mehta and Patel (1983) and is incorporated into the program StatXact. Related but different is the use of multivariate generating functions of Dinwoodie (1998). These can be more efficient than complete enumeration but can also be memory intensive because calculations are done symbolically using commutative algebra with rational coefficients and without round-off simplifications.

The Monte Carlo methods are typically easier to program and are less memory intensive. These can be of various types, such as Markov Chains, rejection methods, and others. Markov chains and rejection methods must be used with great care because of convergence and efficiency problems. The Monte Carlo method we use is one that is perfect in the sense that it produces iid tables with the right hypergeometric distribution, and would be classified as "other" (although it can be viewed as some kind of Markov chain if necessary). So the results are quite reliable (see Table 4 for convergence diagnostics). The paper of Agresti (1992) in describing existing software for exact tests does not say that the Monte Carlo procedure that we use is currently implemented in commercial software, although it has been known for some time and it is very practical. More about this will be said in §3.

Our problem in more detail involves triangular tables with structural zeros along the diagonal and fitting parametric models to such tables, such as the model of quasi-independence first described by Goodman (1968). We are also interested in the goodness-of-fit of more complex parametric models for such tables. Such models are usually analysed using normal approximations (see Bishop, Fienberg and Holland (1977)), which may not be appropriate for sparse high-dimensional data.

Previous work on triangular tables appears in McDonald and Smith (1995). The authors simulate triangular tables with fixed row and column sums using successive conditional one-dimensional hypergeometric distributions to complete the table. This leads to an exact Monte Carlo conditional test for quasi-independence. Their method is not exactly the same as the one we use described in §3 but it is similar in spirit and effect. The simulation method as we describe it goes back to Karl Pearson (see Stigler (1992)) and is described clearly in Diaconis, Graham and Holmes (1999). One feature of the method as described here in §3 is that it can be easily modified for testing goodness-of-fit for an enhanced parametric model that requires simulation on a table that is not even triangular. Finally, a simple symmetric Markov chain can be constructed on these tables, but convergence is relatively slow. The only reason to use a Markov chain would be to obtain the uniform distribution.

Our main interest here is to apply these methods to the following problem from human genetics. Fraser and Hunter (1975) published the following table of pairs of siblings with different types of congenital heart malformations. Only pairs exhibiting different malformations were included as it was easier to collect such data (since it is well known that the same congenital heart malformation often occurs in different siblings). An attempt to provide evidence of non-random association of different defects within families was made by calculating the rank correlation and then doing multiple chi-square tests. Of particular interest was to know whether the malformations ToF and PS were related.

Table 1. Distribution of pairs of siblings with unlike cardiac malformations-major lesion approach from Fraser and Hunter (1975)

	ToF	VSD	PS	TGV	PDA	AS	ASD	Tru	TA	CoA	Dex	Ptr	A − V	Total
ToF	–	13	19	10	4	1	1	0	1	0	1	2	0	52
VSD		–	3	5	3	3	6	1	0	0	2	1	0	24
PS			–	2	0	1	1	3	1	0	0	0	0	8
TGV				–	4	1	2	1	0	1	0	0	0	9
PDA					–	2	0	1	2	0	0	0	1	6
AS						–	2	0	1	3	2	0	0	8
ASD							–	0	1	1	0	0	1	3
Tru								–	0	0	0	1	0	1
TA									–	0	0	0	0	0
CoA										–	0	0	0	0
Dex											–	0	0	0
Ptr												–	0	0
A − V													–	–
Total	0	13	22	17	11	8	12	6	6	5	5	4	2	111

2. Fitted Models.

We consider two parametric models for multinomial probabilities for the triangular Table 1 with vanishing diagonal entries. The first of these is the quasi-independence model, described in Goodman (1968) and Bishop, Fienberg and Holland (1977). This model has a 22-dimensional parameter space that can be described with positive parameters $\alpha_1, \ldots, \alpha_{12}, \beta_2, \ldots, \beta_{13}$ and

$$\begin{gathered} p_{ij} = \alpha_i \beta_j, \quad 1 \le i < j \le 13 \\ \sum_{1 \le i < j \le 13} \alpha_i \beta_j = 1 \\ \beta_1 = 1. \end{gathered} \tag{2.1}$$

This can be put in the form of an exponential family with 22 free parameters $\theta_2, \ldots, \theta_{12}, \gamma_2 = 1, \gamma_3, \ldots, \gamma_{13}$:

$$\begin{aligned} p_{1,j} &= \frac{e^{\gamma_j}}{z_{\theta,\gamma}}, \quad 1 < j \le 13 \\ p_{i,j} &= \frac{e^{\theta_i + \gamma_j}}{z_{\theta,\gamma}}, \quad 1 < i < j \le 13 \end{aligned} \tag{2.2}$$

To go from (2.2) back to (2.1), let $\alpha_1 = 1/z_{\theta,\gamma}$, $\alpha_i = e^{\theta_i}/z_{\theta,\gamma}$ $(1 < i \le 12), \beta_1 = \beta_2 = 1, \beta_j = e^{\gamma_j}$ $(2 < j \le 13)$.

The enhanced model has an additional parameter δ for box $(1,3)$, which in exponential form has 23 free parameters:

$$\begin{aligned}
p_{1,2} &= \frac{1}{z_{\theta,\gamma,\delta}} \\
p_{1,3} &= \frac{e^{\gamma_3+\delta}}{z_{\theta,\gamma,\delta}} \\
p_{1,j} &= \frac{e^{\gamma_j}}{z_{\theta,\gamma,\delta}}, \quad 3 < j \leq 13 \\
p_{i,j} &= \frac{e^{\theta_i+\gamma_j}}{z_{\theta,\gamma,\delta}}, \quad 1 < i < j \leq 13
\end{aligned} \tag{2.3}$$

where the normalizing constant $z_{\theta,\gamma,\delta}$ is the sum of the numerators over the $\binom{13}{2}$ boxes.

The maximum likelihood estimates based on the table (1.1) are below. Each entry in the 13×13 matrix is a triple of numbers, where the first is the data, the second is 111 times the fitted probabilities for quasi-independence (2.1), and the third is 111 times the fitted probabilities for the enhanced model (2.3).

Table 2. Data and Fitted Expected Frequencies for Two Parametric Models

	ToF	VSD	PS	TGV	PDA	AS	ASD	Tru	TA	CoA	Dex	Ptr	A − V	Total
		13	19	10	4	1	1	0	1	0	1	2	0	
ToF	–	13.0	13.6	8.8	4.4	2.7	3.1	1.4	1.2	1.1	1.1	0.9	0.4	52
		13.0	19.0	6.9	3.5	2.1	2.5	1.1	1.1	0.9	0.9	0.7	0.4	
			3	5	3	3	6	1	0	0	2	1	0	
VSD		–	8.4	5.4	2.7	1.7	1.9	0.9	0.8	0.7	0.7	0.5	0.3	24
			3.0	7.3	3.7	2.2	2.6	1.2	1.1	0.9	0.9	0.7	0.4	
				2	0	1	1	3	1	0	0	0	0	
PS			–	2.8	1.4	0.8	1.0	0.4	0.4	0.4	0.4	0.3	0.1	8
				2.8	1.4	0.8	1.0	0.4	0.4	0.4	0.4	0.3	0.1	
					4	1	2	1	0	1	0	0	0	
TGV				–	2.4	1.5	1.7	0.8	0.7	0.6	0.6	0.5	0.2	9
					2.4	1.5	1.7	0.8	0.7	0.6	0.6	0.5	0.2	
						2	0	1	2	0	0	0	1	
PDA					–	1.3	1.6	0.7	0.7	0.5	0.5	0.4	0.2	6
						1.3	1.6	0.7	0.7	0.5	0.5	0.4	0.2	
							2	0	1	3	2	0	0	
AS						–	2.7	1.2	1.1	0.9	0.9	0.8	0.4	8
							2.7	1.2	1.1	0.9	0.9	0.8	0.4	
								0	1	1	0	0	1	
ASD							–	0.7	0.6	0.5	0.5	0.4	0.2	3
								0.7	0.6	0.5	0.5	0.4	0.2	
									0	0	0	1	0	
Tru								–	0.3	0.3	0.3	0.2	0.1	1
									0.3	0.3	0.3	0.2	0.1	
										0	0	0	0	
TA									–	0.0	0.0	0.0	0.0	0
										0.0	0.0	0.0	0.0	
											0	0	0	
CoA										–	0.0	0.0	0.0	0
											0.0	0.0	0.0	
												0	0	
Dex											–	0.0	0.0	0
												0.0	0.0	
													0	
Ptr												–	0.0	0
													0.0	
A − V													–	–
Total	0	13	22	17	11	8	12	6	6	5	5	4	2	111

3. Computation and Analysis.

For computation under the hypergeometric distribution, an exact method of simulation for triangular tables with fixed row and column sums goes back to Karl Pearson (see Stigler (1992)) and is described clearly in Diaconis, Graham, and Holmes (1999). This can be used for the exact (conditional) goodness-of-fit test for the model of quasi-independence. However, for testing goodness-of-fit for the enhanced model, the sufficient statistics are the row and column sums as well as the count in box $(1,3)$. Therefore, sampling must be done with fixed row and column sums and fixed counts of $13, 19, 3$ in boxes $(1,2), (1,3), (2,3)$ respectively. This can be done by modifying the basic full triangular scheme: fill column 3 by drawing 17 balls of Types 1,2,3 from an urn of $52 - 32 = 20$ Type 1, $24 - 3 = 21$ of Type 2, and 8 of Type 3. Then remove this draw, and draw 11 for column 4 from the remaining of Types 1, 2, 3, together with 9 of Type 4 (the row 4 total). These Monte Carlo procedures are much more efficient than Markov chain methods, because they produce independent tables with exactly the right hypergeometric distribution each time. The generating function methods of Dinwoodie (1998) are applicable to this problem, but are much more demanding computationally than the Monte Carlo method.

For the model of quasi-independence with 22 free parameters, the value of the χ^2 goodness-of-fit statistic is 76.1. Using the asymptotic $\chi^2(\binom{13}{2} - 1 - 22 = 55 \texttt{ df})$ distribution, the asymptotic p-value for the goodness-of-fit test is approximately .031. The exact conditional χ^2 test computed with a Monte Carlo method yields a p-value of 0.006, based on a sample of size $100,000$.

For the enhanced model of quasi-independence plus the 23rd parameter δ for box $(1,3)$, the χ^2 goodness-of-fit statistic is 65.9, which on the asymptotic scale of $\chi^2(54)$ gives an asymptotic p-value of 0.13. The Monte Carlo exact method gives a p-value of .036.

To test the significance of the 23rd parameter δ under a one-sided test

$$H_0 : \delta = 0$$
$$H_1 : \delta > 0$$

the exact Monte Carlo simulation found a p-value of .003. This number is the conditional probability of a count of 19 or more in box $(1,3)$ with respect to the hypergeometric distribution on triangular tables (with vanishing diagonal) with fixed row and column sums equal to those of the observed data.

The following table of exact p-values summarizes the analysis.

Table 3. Exact conditional p-values

Test	$p -$ value
Quasi $-$ independence fit	0.006
Enhanced Model fit $(+\delta)$	.036
$H_0 : \ \delta = 0$	.003

To explain in more detail the sampling for the enhanced model with δ, consider the underlying problem of sampling from the hypergeometric distribution from tables with fixed row and column sums of the form:

$$\begin{array}{ccccc}
n_{13} & n_{14} & n_{15} & n_{16} & r_1 \\
n_{23} & n_{24} & n_{25} & n_{26} & r_2 \\
n_{33} & n_{34} & n_{35} & n_{36} & r_3 \\
- & n_{44} & n_{45} & n_{46} & r_4 \\
- & - & n_{55} & n_{56} & r_5 \\
- & - & - & n_{66} & r_6 \\
c_3 & c_4 & c_5 & c_6 & n
\end{array}$$

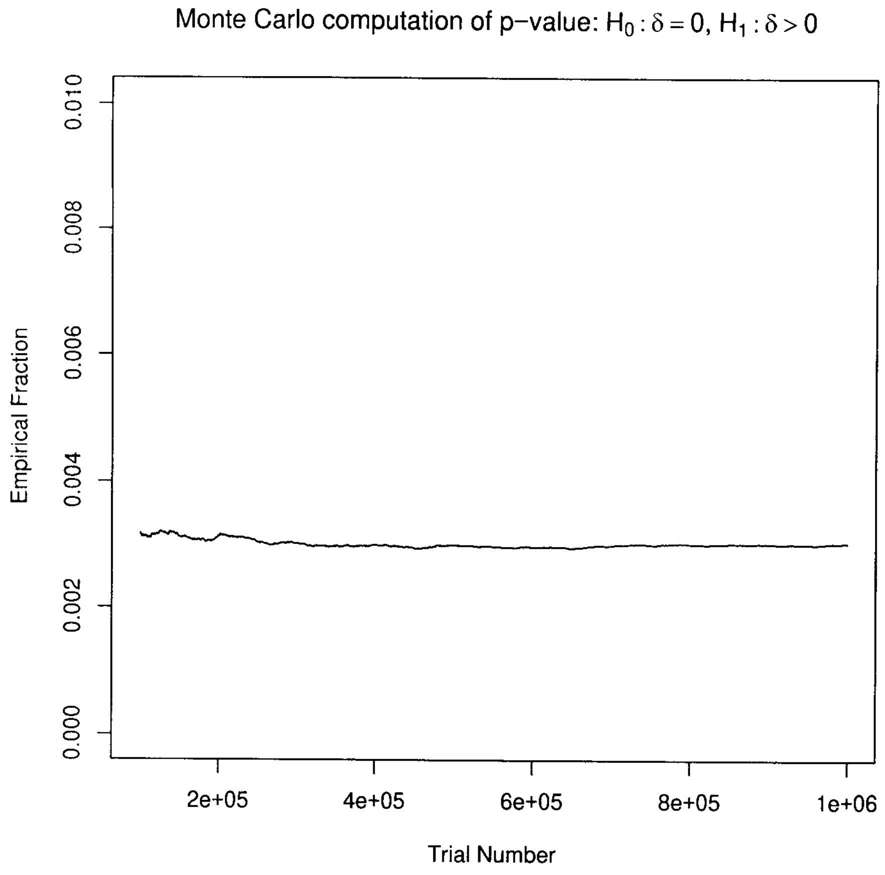
Suppose there are r_1 balls of color R_1, r_2 of color R_2, etc., and view the values $c_3, c_4, \ldots, c_6$ as sample sizes. Sample c_3 from the colors R_1, R_2, R_3, and remove the result. Then sample c_4 from the remaining first 3 colors and also R_4, etc. This results in the factorization of the hypergeometric as follows:

$$\begin{aligned}
P(\mathbf{n}) = & \frac{\binom{r_1}{n_{13}}\binom{r_2}{n_{23}}\binom{r_3}{n_{33}}}{\binom{r_1+r_2+r_3}{c_3}} \\
& \times \frac{\binom{r_1-n_{13}}{n_{14}}\binom{r_2-n_{23}}{n_{24}}\binom{r_3-n_{33}}{n_{34}}\binom{r_4}{n_{44}}}{\binom{r_1+r_2+r_3+r_4-c3}{c_4}} \\
& \times \frac{\binom{r_1-n_{13}-n_{14}}{n_{15}}\binom{r_2-n_{23}-n_{24}}{n_{25}}\binom{r_3-n_{33}-n_{34}}{n_{35}}\binom{r_4-n_{44}}{n_{45}}\binom{r_5}{n_{55}}}{\binom{r_1+r_2+r_3+r_4+r_5-c_3-c_4}{c_5}} \\
& \times \frac{\binom{r_1-n_{13}-n_{14}-n_{15}}{n_{16}}\binom{r_2-n_{23}-n_{24}-n_{25}}{n_{26}}\binom{r_3-n_{33}-n_{34}-n_{35}}{n_{36}}\binom{r_4-n_{44}-n_{45}}{n_{46}}\binom{r_5-n_{55}}{n_{56}}\binom{r_6}{n_{66}}}{\binom{r_1+r_2+r_3+r_4+r_5+r_6-c_3-c_4-c_5}{c_6}}
\end{aligned}$$

which is the hypergeometric distribution in the variables n_{ij}. This sampling scheme was implemented on full-size tables $(c_3, \ldots, c_{12})$ using the `sample` command in the language R for multivariate sampling without repetition to simulate tables.

To see convergence of the empirical fractions with a sample of 100,000, we plotted the empirical fraction estimates of the p-value versus the trial number. Below is the graph for the computation for testing $\delta = 0$.

Table 4. Diagnostic Ouput



4. Discussion and Conclusion.

We have employed Monte Carlo methods of exact inference to complete the statistical analysis of a genetic study of sibling pair data that had been studied previously by Fraser and Hunter (1975) and MacGibbon (1983). The data is in the form of a triangular table with vanishing diagonal entries. Simulation required a slight extension of a known efficient method for triangular tables. The exact analysis confirms and refines earlier conclusions.

Acknowledgments. This work was supported by NSERC of Canada, and the State of Louisiana. We thank GERAD and ROHCP of the Université de Montréal for the use of computing facilities. We used the software R from lib.stat.cmu.edu for simulations. We also thank John McDonald for related references.

References

1. A. Agresti (1992). A survey of exact inference for contingency tables, *Statistical Science,* **7**, 131-177.

2. A. Agresti (1999). Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine*, to appear.

3. Y. N. M. Bishop, S. E. Fienberg, P. W. Holland (1977). *Discrete Multivariate Analysis.* The MIT Press, Cambridge MA.

4. P. Diaconis, R. Graham, S. P. Holmes (1999). Statistical problems involving permutations with restricted positions. *Festschrift in Honor of William Van Zwet*, to appear.

5. I. H. Dinwoodie (1998). Generating functions for exact p-values of odds ratios in logistic regression. *J. Ital. Statist. Soc.*, **3**, 221-232.

6. F. C. Fraser and A. D. W. Hunter (1975). Etiologic relations among categories of congenital heart malformations. *The American Journal of Cardiology*, **36**, 793-796.

7. L. A. Goodman (1968). The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with or without missing entries. *JASA*, **63** 1091-1131.

8. B. MacGibbon (1983). A log-linear model of a paired sibling study. In *Proceedings of Statistics '81 Canada Conference,* eds. Y. Chaubey, T. D. Dwivedi, 193-197.

9. J. W. McDonald and P. W. F. Smith (1995). Exact conditional tests of quasi-independence for triangular contingency tables: estimating significance levels. *Applied Statistics*, **44**, 143-151.

10. P. W. F. Smith and J. W. McDonald (1994). Simulate and reject Monte Carlo exact conditional tests for quasi-independence. In *Proceedings of COMPSTAT 1994,* eds. Dutter, R. and W. Grossman, Physica-Verlag, Heidelberg, 509-514.

11. S. Stigler (1992). Studies in the history of probability and statistics XLIII. Karl Pearson and quasi-independence. *Biometrika*, **79**, 563-575.